The Effect of College Attainment on Per-Capita Income across U.S States

Noa Tshimanga

Abstract

College attainment is often cited as one of the main explanatory variables that can explain the variation in per-capita income. This study uses cross-sectional data to analyze the relationship between the education and income using additional variables that include high school attainment, manufacturing as a percentage of the labor force, expenditure per pupil by state governments as well as state GDP. Most studies centered on this topic usually have focused on trying to examine how inequality in education affects wage. This study takes a different approach as it tries to use conclusive experimental data to try to conclude the foundational relationship between college attainment and per-capita income across states. Using data largely from Bureau of Economic Analysis in conjunction with data from sources such as the U.S Department of Agricultural Services, National Center for Education Statistics, and more, I was able to conclude that college attainment on average across all states does in fact positively correlate with per-capita income and the relationship tends to get stronger as more explanatory variables that correlate with college attainment are added into consideration.

I. Introduction

The importance of education is not something of a new phenomenon in the United States of America. In the early founding years of America, education was regarded as sacred even though it was mostly for those of elite status. But as America evolved from a cluster of 13 colonies to a gradually large nation, there started to be a change in the access of education. From the first and oldest public school in Boston, Massachusetts, education in America shifted from being possessed only by the rich and elite to the common population (The American Board, 2015). Public schools began to be funded through district lines drawn by law representatives; households funded education through the pooling of resources, in this case, specifically property taxes (Gershon, 2016); and it even became mandatory by the 19th century for all children to attend schools until a certain age had been reached (The American Board, 2015). This is because the benefits of education outweighed the cost to the common public. By 1940, the high school graduation rate in America increased from 9% in 1910 to 51%, known as the "High School Movement" (Goldin & Katz, 2008), this period was identified as a grand marker in the expansion of education in America. As a consequence, as education continued to grow, higher education became even more important. There started to be an emphasis on higher education in achieving a desirable stable salary in society. As a result, the United States has been experiencing an upward trend in the percentage of individuals attaining a college education due to the value and return in investment on higher education.

Current studies show that a new year of education has the capability to increase wages between 8 and 13% (FRED, 2017). Of course, these effects can differ based on state of employment, gender, race, and numerous other factors. But the fact remains that studies have discovered some sort of causal relationship between education and money. In a study done by Berkley, it was estimated that by 2020, 65% of all jobs in the economy will require a postsecondary education and training beyond high school. In measuring the value of education, it is clear that as individuals transition from primary to secondary and to higher education the median salary increases as respective unemployment rates decrease. Therefore, in terms of dollars, higher education makes perfect sense.

This paper strives to draw a more conclusive relationship between educational attainment and income using cross-sectional data. In particular, this paper will examine and develop the relationship between college attainment and individual income across the 50 states, including the District of Columbia. My hypothesis is that there will be a positive correlation between college attainment and per-capita income across states. This hypothesis was generated on the rationale that as more states

increase and develop their human capital (college attainment) then the return should be seen on an individual level regardless of the existing differences that can exist across states.

II. Literature Review

Houthakker (1959) study draws on the relationship between school attendance in years completed and money gained to examine the relationship between income and education. Using a cross-sectional data set, this paper offers a contribution to the study of education and income by breaking down years of education into subcategories of age brackets as well breaking down income into brackets represented by amounts that are before and after tax. The most relevant piece of this literature is its finding on how education leads to a higher median of income. An interesting detail of this result is that this trend is seen with individuals under 30 rather than across all age brackets indicating that the marginal return education might decrease as individuals transition through age brackets. Houthakker(1959) also touches on the inequalities that exist within the access of education, citing that individuals who come from high income families will most likely be those who have access to education and therefore are the ones that will experience a higher median income with increasing years of education. It is important to note the year in which this piece of literature was written as the access to education was drastically different in 1959 compared to 2019, the year this cross-sectional data paper focuses on.

Gregorio and Lee (1999) discuss the effect of education on income distribution. This paper hypothesizes that higher educational attainment leads to more equal income distribution across countries. The study runs a regression using the gini coefficient constructed from raw data obtained from using regional dummies to account for differences of income across countries. In running the regression, it was expected to reveal that the gini coefficient was to be greater in the analysis of countries that provide more individual data. This literature also makes the distinction of income before and after tax to control for the differences in tax systems across countries. The results from the regression lead to the interpretation that higher educational attainment leads to higher income equality and thereby highlights the results that as attainment of education continues to grow, this leads to less dispersion of income. Gregorio and Lee (1999) also find that a 10% increase in per capita income increases educational attainment by .01 years indicating that as income grows, educational attainment grows. This is relevant to this paper in that it provides evidence of the relationship between per capita income and educational attainment. Though the relationship analyzed in this specific literature is reversed from my focus of analysis, it provides context in the way the data, namely how it explains the relationship between per-capita income and educational attainment, from this paper can be further analyzed. A difference to

note is that while the previous relevant literature provided units of observation focused on households, Gregorio and Lee (1999) provide analysis on countries of sample size of 90 using panel data.

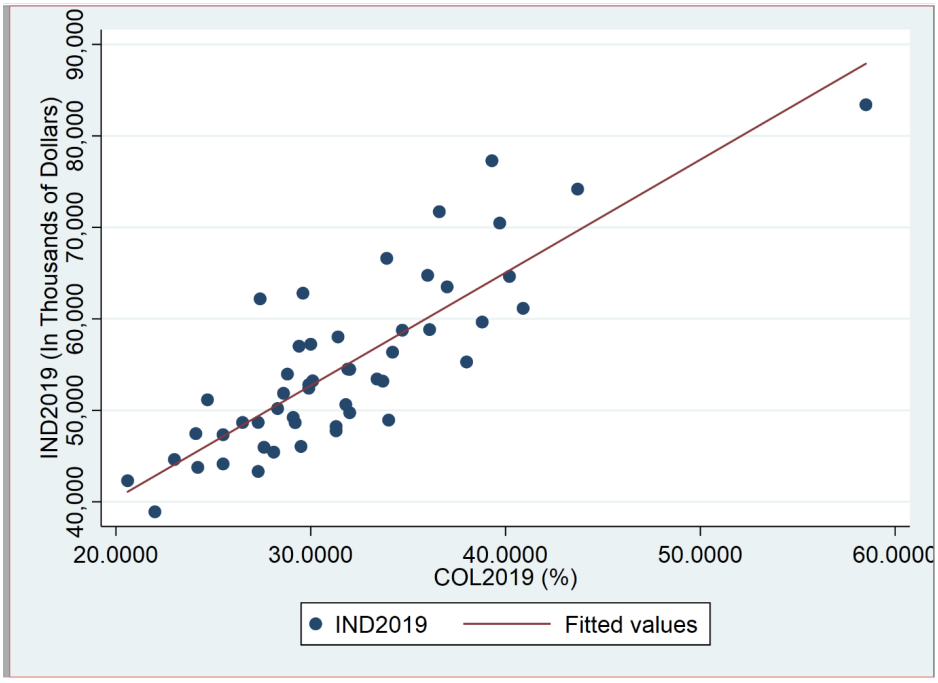
Turner (2007) takes a different new approach in analyzing the relationship between education and income. It is notable that this specific makes two major contributions in that it introduces original annual years of school and measures the average years of the experience in the labor force for each state. This piece of literature makes use of 5 decades worth of U.S census data and further estimates data from an additional 5. To characterize the relationship between education and income, Turner (2007) uses a perpetual inventory method, employed by Barro and Lee (1993) and Baier, Dwyer, and Tamura (2006), to construct average years of schooling in the labor force for each state. This is important in that other relevant literature will construct average years of schooling for state residents rather than just those who are a part of the labor force. The relationship between educational attainment and income is then defined using a new angle of analysis. In the paper, Turner (2007) Takes into account migration and different levels of education such as primary, secondary, and tertiary, and in addition variables that can affect the labor force. States are also grouped up in their respective regions to offer a more in depth analysis on how education affects per capita income for those in the labor force. Last, but not least, variance of educational attainment across states is taken to allow for differences across collected observations. Results lead to the assumption that higher levels of education do correlate with positive returns on education for those in the labor force.

It is evident that it is not a new phenomenon to study the impact of education on income. There have been many different approaches to this study, each with a new question to answer. Therefore though this paper takes on many underlying foundational tactics utilized in the relevant literature discussed and beyond, it still makes a contribution to the existing literature. Most of the primary literature found during the process of research focused on taking the relationship between income and education to develop a hypothesis on income inequality and therefore educational inequality. What differentiates this paper is that it takes a more general approach and does not use a pre-existing relationship to further develop a new one. This paper makes a contribution to the existing literature by solely focusing on states and the percentage of educational attainment that each state has obtained and how that impacts per-capita income rather than using individuals, multiple countries, or households as the units of observations. In addition, this paper makes a contribution to the existing literature in the explanatory variables included. While the main explanatory variable of this paper is common, accounting for the percentage manufacturing jobs in the labor force designates a new insight into the relationship between individual income and the percentage of college attainment across U.S states.

III. Data

To analyze and develop the relationship between individual income and educational attainment, cross-sectional data of each variable was gathered in the year 2019. Individual income of each of the 50 states in addition to the District of Columbia was gathered from the Bureau of Economic Analysis from the annual report of per capita income. The primary explanatory variable is educational attainment represented by the percentage of the state population to earn a college degree, whether it be a graduate or undergraduate degree. Data for the primary explanatory variable was taken from the U.S Department of Agriculture, Economic research service. This explanatory variable was picked due to the already existing causal impact that education has on income in general. There is no need to differentiate between the percentage of those with an undergraduate degree and those with graduate degrees because the focus of the study is not to determine a causal relationship between increasing amounts of education and income. Instead, the explanatory variable captures an overall general relationship of how the highest educational attainment affects individual income across states, again including the District of Columbia. Figure 1. below is a scatter plot of the primary explanatory variable, COL2019 and the explained variable, IND2019. As expected, there lies a positive correlation between percentage of educational attainment in the population and per-capita income across states.

Figure 1. Scatter plot of IND2019 vs. COL2019



4 other variables are included beyond the main two variables discussed in order to reveal the ceteris paribus relationship between individual income and percentage of educational attainment and create a fitting multiple linear regression model: percentage of population to earn a high school diploma, represented by HS2019 is included in the model, expenditure on schooling per pupil, EXDPUP2017, percentage of jobs within the workforce belonging to manufacturing, MAN2019, and state GDP, STATEGDP2019. Percentage of population to earn a high school diploma and COL2019 can be highly correlated. It is to differentiate between states with differing levels of college and high school graduation and how that might in turn affect individual income across states. Data for HS2019 was sourced from the U.S Department of Agriculture, Economic Research service. It is hypothesized that within the multiple linear regression model that its coefficient will carry a positive sign indicating a positive correlation with individual income, as states with higher levels of high school education attainment levels would be more likely to experience higher levels of per capita income in thousands of dollars. The third explanatory variable, represented by EXDPUP2017 explains how much in thousands of dollars states spend per student on education. It is hypothesized that states with a high per pupil expenditures on education will experience higher per-capita income, as investment in education tends to lead to positive returns represented by income. A minor discrepancy is important to note for the explanatory variable of expenditures per pupil. While the rest of the explanatory and explained variables come from 2019 cross-sectional data, EXDPUP comes from 2017 data. This is due to the lack of available data from the National Center of Educational Statistics, where expenditure per pupil was sourced from. Reports start from the school year of 1960 until 2017. This paper proceeds with 2017 data because there is no belief that there is a significant difference between expenditure per pupil in 2017 and in 2019. Next, MAN2019 represents the percentage of jobs in the workforce that belongs to the manufacturing industry. This explanatory variable differs importantly from the rest because it is hypothesized to have negative correlation with the explained variable and the primary explanatory variable. This is because average to low-level manufacturing jobs tend to have lower levels of education, more commonly, a minimum requirement of high school diploma. The hypothesis carries that states that have higher percentages of manufacturing jobs will tend to have less levels of per-capita income. Data for MAN2019 was taken from the National Association of Manufacturers. The last explanatory variable is that state GDP. It is hypothesized that the this variable will carry a positive sign as the higher states earn, the higher their residents should earn. This data was extracted from the Bureau of Economic Analysis.

Table 1. Summary of Variables 1.1

Variable Name	Description	Year	Units	Source
IND2019	Income per capita	2019	Thousands of Dollars	Bureau of Economic Analysis
COL2019	Percentage of population with college education attainment	2019	Years of education (at least 4)	U.S Department of Agricultural Services
HS2019	Percentage of population with high school education attainment	2019	Years	U.S Department of Agricultural Services
EXDPUP2017	Expenditure of Education per pupil	2017	Thousands of Dollars	National Center for Education Statistics
MAN2019	Percentage of manufacturing jobs a part of the workforce	2019	Thousands of jobs	National Association of Manufacturers
STATEGDP2019	Gross Domestic Product per state	2019	Millions of dollars	Bureau of Economic Analysis
LogIND2019	Log of income per capita	2019		Bureau of Economic Analysis
logEXD2019	Log of income per capita	2019		National Center of Economic Analysis
CST2020	Composite cost of living per state	2020	Index; US = base = 100	Missouri Economic Research and Information Center
logSTATEGDP2019	Log of Gross Domestic Product per state	2019		Bureau of Economic Analysis
URBAN2010	Percentage of population in urban areas	2010	Percentage of total population	United States Census Bureau

Table 2. Summary of Variables 1.2

Variable	Observation	Mean	Std. Dev.	Min.	Max.
IND2019	51	54806.49	9471.80	38914	83406
COL2019	51	31.70	6.43	20.6	58.5
HS2019	51	28.01	4.17	16.8	40.3
EXDPUP2017	51	13185.22	4004.37	7521	23861
MAN2019	51	8.27	3.67	1.42	17.07
logIND2019	51	10.89789	.1646007	10.56911	11.33148
logEXD2017	51	9.445505	.2853371	8.925454	10.08
CST2020	51	105.7431	21.52505	84.5	198.6
STATEGDP2019	51	371142.9	488423.3	29806.2	2800506
logSTATEGDP2019	51	12.26834	1.05195	10.30247	14.84531
URBAN2010	51	74.0998	14.88518	38.66	100

The data is assumed to meet all 5 Gauss Markov Assumptions.

1. **The model is to be linear in parameter: $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_5 + u$**. All parameters of the relevant model are linear in parameters therefore the assumption is met.
2. **Random Sampling: Data is to be collected by measures of random sampling.**
   Data was collected for all 50 states plus the District of Columbia from all available sources, reaching the maximum number of observations indicating no room for manipulation of the gathering data. Therefore, data meets the second assumption of random sampling and satisfies the second Gauss Markov assumption

3. **No Perfect Collinearity: No perfect correlation exists between the explanatory variables and no explanatory variables are constants**. None of the explanatory variables are constants and perfectly correlated. This assumption is further proved by STATA analysis.
4. **Zero Conditional Mean: The error term, u has a value of 0 for any value of the independent variable.** This assumption is assumed to be met through the assumption that the model is not misspecified and that there is no correlation between any explanatory variables and u.

Because the data is assumed to meet the Gauss-Markov Assumptions 1-4, it is assumed to be unbiased.

5. **Homoskedasticity: The error term, u has the same variance given any value of the explanatory variable.** In other words, u is a constant. This last assumption, though not required to prove unbiasedness, is important nonetheless to show that the variance of u does not depend on the values of the explanatory variables. It is assumed in this model, variance of u does not depend on COL2019, the main explanatory variable or any other other explanatory variables in the model. Therefore, assumption 5 is met.

The Classical Linear Model Assumptions

This 6th assumption in conjunction with the Gauss-Markov assumptions make up the Classical Linear Model Assumptions and the data is assumed to meet assumptions 1-6 and therefore the Gauss-Markov assumptions as well as the Classical Linear Model Assumptions.

6. **Normality: The population error, u, is independent of the explanatory variables and is normally distributed with zero mean and variance: u ~Normal (0, σ^2).** This assumption allows the sampling distribution of β_j-hat to be tractable by assuming that u, the error term, is normally distributed. By definition, when making this assumption, assumptions 4 & 5 from above are also naturally assumed. This model of assumption is used to show that OLS estimators are the minimum variance unbiased estimators. In addition, when summarizing the population assumptions of the CLM assumptions, it is found that y conditional on x has a normal distribution with mean linear in explanatory variables and a constant variance. Because data is assumed to have an error term, u, that has a normal distribution, assumption 6 is met.

IV. Results

1. **Model 1. The Simple Linear Regression Model**

The model equation is as follows:

$$\mathbf{IND2019 = \beta_0 + \beta_1 (COL2019) + u}$$

The estimated equation with a sample size of 50 states and 1 district, the District of Columbia, is as follows:

IND2019 =15671.61 +1234.62(COL2019)

(3703.589) **(**114.55)

n = 51, R^2 = 0.70

By first starting out to interpret the coefficient of the main explanatory variable in the simple regression model, the coefficient of COL2019 is seen to have a positive sign as hypothesized indicating that a 1 percentage point increase in the college attainment across states can, on average, increase individual income by 1234.64 dollars. Examining the R^2 value, we see that even in the simple linear regression model that COL2019 accounts for an explanation of 70.00% of per-capita income variation across states indicating that COL2019 does indeed have a significant impact on per-capita income. In addition, COL2019 and the intercept was found to be statistically significant at the 1% level. But while this simple regression is a good start to examining the relationship between per-capita income and college attainment, it does not capture the full story.

2. Model 2. The Multiple Linear Regression Model

The second model is represented by the multiple linear regression model of the explained variable, the main explanatory variable, and additional explanatory variables to capture the ceteris paribus relationship between IND2019 and COL2019.

The model equation is as follows:

$IND2019 = \beta_0 + \beta_1 (COL2019) + \beta_2(HS2019) + \beta_3(EXDPUP2017) + \beta_4(MAN2019) + \beta_5(STATEGDP2019) + u$

The estimated equation with a sample size of 50 states and 1 district, the District of Columbia, is as follows:

IND2019 = 3178.89 + 519.65(COL2019) - 362.03(HS2019) + 1.25(EXDPUP2017) - 87.18(MAN2019)

(10030.16) (181.31) (240.03) (.21) (164.46)

+ .003(STATEGDP2019)

n = 51 R^2 = 0.87 (.001)

What is important to note is the difference in magnitude of the coefficient for COL2019 in the simple linear regression and the multiple linear regression. In the simple linear regression, the estimated β_1 coefficient is 1234.62 while in the multiple linear regression, the estimated β_1 coefficient decreases to 519.65. When a simple regression of HS2019, X_2 on COL2019, X_1 is performed, we obtain a δ value of -.48, with the following estimated simple regression equation: HS2019 = 43.09 -.48 (COL2019)

This indicates that the bias of β_1 when omitting X_2 is positive by the relation of the sign of β_2 in the multiple linear regression model, which is negative, multiplied by the sign of the correlation between HS2019 and COL2019, which is negative, as indicated in the estimated simple regression equation of HS2019 on COL2019 above.

What turns out to be surprising about the results of the multiple linear regression analysis is that the coefficient for HS2019 carries a negative sign indicating a negative correlation with the IND2019 which is the opposite of what was previously hypothesized. It can be interpreted that a one percentage point increase in high school attainment across states leads to a decrease of 362.03. Because this specific variable describes those who have *only* attained a high school diploma and do not move on further to higher education, it can be assumed that if states experience increasing amounts of individuals whose highest educational attainment is only a high school diploma then that would on average have a negative impact on individual income across states.

EXDPUP2017 and MAN2019 both exhibit smaller ratios of coefficients compared to the previous two discussed. The coefficient for EXDPUP2017 is positive as hypothesized and equals 1.35 while the coefficient for MAN2019 is negative and equals -2.36. Though having small coefficient, state GDP also carries a positive sign as hypothesized, but the magnitude of the coefficient brings doubt to the interpretation of the effect of state GDP on per-capita income, all else equal.

Last, in the multiple linear regression the R^2 value increases to explaining 87% of per-capita income variation across states indicating that this model is better fit than its simple linear regression counterpart.

3. Model 3. Statistically Significant (I)

The third model is represented by regressing the dependent variable on the explanatory variables that were found to be statistically significant during the conduction of the t-test in model 2.

The model equation is as follows:

$$\mathbf{IND2019 = \beta_0 + \beta_1 (COL2019) + \beta_2 (EXDPUP2017) + \beta_3 (STATEGDP2019) + u}$$

The estimated equation with a sample size of 50 states and 1 district, the District of Columbia, is as follows:

$$\mathbf{IND2019 = 14831.71 + 764.86 (COL2019) + 1.1 (EXDPUP2017) + .003 (STATEGDP2019)}$$

(2632.13) (106.40) (0.17) (0.001)

$n = 51$ $R^2 = 0.86$

This model drops the variables, HS2019, and MAN2019 due to being statistically insignificant in model 2. The first two things to notice is that there is not a large difference between the coefficients in model 2

and model 3 for the EXDPUP2017 and STATEGDP2019, in fact the coefficient and the standard error of STATEGDP2019 is exactly the same when rounded to three decimal places; even in this model with only statistically significant explanatory variables, their coefficients remain relatively small. Nevertheless, even with small coefficients, both variables still remain statistically significant in this estimated model, implying that a one thousand dollar increase in expenditure per pupil leads to a 1.1 increase in per-capita income and a one million dollar increase in GDP is correlated with .003 increase in per-capita income, holding all else equal. The intercept of this estimated model is rather large compared to that of the model 2. This is the same for the coefficient of COL2019. This estimated model implies that a one percentage point increase in COL2019 leads to an increase of 764.86 in individual income. The R-squared value of this model remains relatively high at a value of .86. Compared to model 2, there was a decrease by .01.

4. **Model 4. Cost of Living and Urban Population**

This fourth model is a multiple linear regression that includes all the original explanatory variables included in model 2, but specifically adds in two additional independent variables: CST2020 and URBAN2010, an index representing the cost of living in each state in 2020 and the percentage of the population that lives in an urban area, respectively. These two variables were not added into the main model because of the years in which the data was published. As a result, caution is to be taken with this model as large assumptions are to be made that the data is not significantly different compared to the focus year of 2019. Nevertheless, this model is still interesting because cost of living and percentage of urban population can be thought to correlate with per-capita income.

The model equation is as follows:

$$\mathbf{IND2019 = \beta_0 + \beta_1(COL2019) + \beta_2(HS2019) + \beta_3(EXDPUP2017) + \beta_4(MAN2019) + \beta_5(STATEGDP2019) + \beta_6(CST2020) + \beta_7(URBAN2010) + u}$$

The estimated equation is as follows:

$$\mathbf{IND2019 = 22134.52 + 465.15(COL2019) - 270.80(HS2019) + 1.28(EXDPUP2017) - 17.31(MAN2019)}$$
$$(11400.46) \quad (179.71) \quad (242.65) \quad (0.23) \quad (171.10)$$
$$\mathbf{+ .002(STATEGDP2019) + 4.85(CST2020) + 98.88(URBAN2010)}$$
$$(.001) \quad (34.70) \quad (46.95)$$
$$n = 51 \; R^2 = 0 .88$$

By first starting out to examine the variables of interest in this model: CST2020 and URBAN2010, it can be seen that CTS2020 has a relatively small coefficient and during the conduction of the t-test was found to be insignificant. Therefore, though the cost of living in each state does positively correlate with

per-capita income, it does not prove to be a significant correlation with per-capita income, holding all else equal. URBAN2010 holds a different story as a 1 percentage point increase in URBAN2010 leads to a 98.88 increase in individual income and after the conduction of a t-test is found to be statistically significant at the 5% level implying that the percentage of peoples living in an urban population correlates significantly with individual income. COL2019 and EXDPUP2017 remains statistically significant and holds a positive correlation with IND2019 in this model as well though the estimated coefficient attached to COL2019 is smaller than that compared to models 1-3 while the coefficient of EXDPUP2017 is larger though not by much compared to models 1-3. Last, HS2019 and MAN2019 once again fail to be statistically significant but hold the same sign compared to models 1-3. STATEGDP2019, holding all else equal differs from its previous usual behavior and instead is found to be statistically insignificant. In addition, though not a large number by far, its coefficient decreased by .001. Though not both variables of interest turn out to be significant as initially hypothesized, the overall estimated model does have a R-squared value of .88, the largest value thus far, meaning that about 88% of the variation of per-capita income is explained in this model, but caution should should be held because when examining the adjusted R-squared value, it is only .86 exposing the penalty of adding more explanatory variables.

5. **Model 5. The log-log Model**

The fifthmodel is represented by taking the log of IND2019, EXDPUP2017, and STATEGDP2019 to present another way of interpreting the relationship between IND2019 and COL2019 as most of the explanatory variables in the model are already in terms of percentages rather than raw numbers, in attempt to make the model easier to relate and understand.

The model equation is as follows:

$$\mathbf{logIND2019 = \beta_0 + \beta_1(COL2019) + \beta_2(HS2019) + \beta_3(logEXDPUP2017) + \beta_4(MAN2019) + \beta_5(logSTATEGDP2019) + u}$$

The estimated equation with a sample size of 50 states and 1 district, the District of Columbia, is as follows:

$$\mathbf{logIND2019 = 7.45 + .006(COL2019) - 0.006(HS2019) + 0.34(logEXDPUP2017) - 0.001(MAN2019) + 0.02(logSTATEGDP2019)}$$

(0.41) (.003) (.004) (.05) (.003) (.009)

$n = 51$ $R^2 = 0.86$

The first thing to mention is because the independent variables: COL2019, HS2019, and MAN2019 were already in terms of percentages, when the log of model 2 was taken, their coefficients became smaller compared to models 1-4. The variables of interest to interpret in this model, holding all else equal, are logEXDPUP2017 and logSTATEGDP2019. Both variables are statistically significant with logEXDPUP2017

being significant at the 1% level and logSTATEGDP2019 being significant at the 5% level. Both variables still hold a positive correlation, except now it is with logIND2019 rather than IND2019. This implies that a one percent increase in expenditure per pupil leads to a 0.34% increase in per-capita income and a one percent increase in state GDP leads to a 0.02% increase in per-capita income. These results are more simplified and easier to understand compared to the above estimated models showing the influence of expenditure per pupil and the effect of state GDP on individual income. This is also the first model that has shown HS2019 to be statistically significant at the 5% level despite its small coefficient. In addition, the R-squared value for this model remains at 0.86.

6. Model 6.

Model 6 is represented by regressing the dependent variable on the explanatory variables that were found to be statistically significant during the conduction of the t-test in model 5.

The model equation is as follows:

$$\mathbf{logIND2019 = \beta_0 + \beta_1 (COL2019) + \beta_2 (HS2019) + \beta_3 (logEXDPUP2017) + \beta_4 (logSTATEGDP2019) + u}$$

The estimated equation is as follows:

$$\mathbf{logIND2019 = 7.42 + .006\,(COL2019) - .01(HS2019) + 0.35(logEXDPUP2017) + 0.02(logSTATEGDP2019)}$$

(0.40) (.003) (.004) (.05) (.009)

$n = 51$ $R^2 = 0.86$

This model only drops the explanatory variable, MAN2019 due to the fact that it was statistically insignificant in the estimated model 5. What is again interesting to mention about this estimated model is that HS2019 is once again statistically significant now at the 1% level implying that a one percentage point increase in the percentage of the population that only obtain high school diplomas decreases per-capita income by 0.01%. The rest of the explanatory variables demonstrate the same behavior as in model 5 with coefficient values that are close in value and magnitude with differences due to rounding. The coefficient of COL2019 exhibited the same value from model 5 implying that a one percentage point increase in college attainment increases individual income by 0.006%. The coefficients of logEXDPUP2017 and logSTATEGDP2019 also exhibited the same value from model 5, in addition with the same standard error. The R-squared value also remained the same at the value of 0.86.

Table 3. Summary of Regression Models

	Dependent Variable: IND2019				Dependent Variable: logIND2019	
Independent Variables	MODEL 1	MODEL 2	MODEL 3 (model 2)	MODEL 4	MODEL 5	MODEL 6
COL2019	1234.62*** (114.55)	519.65*** (181.31)	764.86*** (106.40)	465.15** (179.71)	.006** (.003)	.006* (.003)
HS2019		-362.03 (240.03)		-270.80 (242.65)	-.009** (.004)	-.01*** (.004)
EXDPUP2017		1.25*** (.21)	1.1*** (.17)	1.28*** (.23)		
MAN2019		-87.18 (164.46)		-17.31 (171.10)	-.003 (.003)	
STATEGDP2019		.003** (.001)	.003*** (.001)	.002 (.001)		
logEXDPUP2017					.34*** (.05)	.35*** (.05)
logSTATEGDP2019					.02** (.009)	.02*** (.009)
CST2020				4.85 (34.70)		
URBAN2010				98.88** (46.95)		
Intercept	15671.61*** (3703.589)	3178.89*** (10030.16)	14831.71*** (2632.13)	22134.52* (11400.46)	7.50*** (.41)	7.42*** (.40)
No. of Obs.	51	51	51	51	51	51
R-squared	0.70	0.87	0.86	0.88	0.86	.86
Adjusted R-squared	0.70	0.85	0.85	0.86	0.84	0.85

IV. Extensions

1. The F-test

Model 3 was created from dropping the variables that were individually found to be statistically insignificant. After producing the correlation chart for the explanatory variables, it was discovered that the explanatory variables were highly correlated, giving concern to the issue of multicollinearity with relatively high standard errors and consequently low t-statistics. Model 3 is therefore the restricted model of model 2 but to find out if these variables should have been dropped from model 2, the main model, to create a restricted model that better explains the variation of per capita income, a F-test was conducted to discover whether or not the variables dropped, HS2019 and Man2019, were jointly significant.

The null hypothesis is as follows:

$$\mathbf{H_0 = \beta_2 = 0, \beta_4 = 0}$$

The alternate hypothesis is as follows:

$$\mathbf{H_1 = H_0 \text{ is not true.}}$$

The F statistic was collected by using the R-squared values of the unrestricted model and the restricted model as well as the number of restrictions, *q*, and the degrees of freedom, (n - k -1). The value generated was 1.78. At the 10% level, the critical value $F_{2,45}$ based on the number of restrictions, q, and degrees of freedom respectively was discovered to be 2.43. The critical value of the F-distribution is greater than the F-ratio, thus we fail to reject the null hypothesis and consequently drop HS2019 and MAN2019 from model 2 as they are jointly insignificant and instead rearrange focus on model 3.

A F-test was also conducted for model 5 out of curiosity as once the functional form of the model 2 was changed to log-log, HS2019 became statistically significant at 5% level while MAN2019 remained statistically insignificant. The main question is after the conduction of a F-test if HS2019 and MAN2019 would remain jointly statistically insignificant or if there would be a change and the null hypothesis would have to be rejected.

The null hypothesis is as follows:

$$\mathbf{H_0 = \beta_2 = 0, \beta_4 = 0}$$

The alternate hypothesis is as follows:

$$\mathbf{H_1 = H_0 \text{ is not true}}$$

Once again, the F-statistic was generated using the R-squared values from the unrestricted model, model 5 and the restricted model that was estimated once dropping HS2019 and MAN2019. The value of the

F-ratio is 2.02. At the 10% level, the critical value $F_{2,45}$ was 2.43. Because once more the critical value is greater than the F-statistic, we fail to reject the null hypothesis and consequently rather only dropping MAN2019 as was done for model 6, both HS2019 and MAN2019 might need to be dropped from model 5 as they are discovered to be jointly insignificant, holding all else equal, regardless of the fact that HS2019 was found to be statistically significant in model 5. Nevertheless hesitation exists as the difference in value of the F-ratio and critical value for the f-distribution was very small.

2. Functional Form

A functional form of the main model is model 5, where the model 2 was transformed into a log-log model. The new dependent variable is logIND2019 and the explanatory variables are COL2019, HS2019, logEXDPUP2017, MAN2019, and logSTATEGDP2019. This log-log form allows us to examine the individual income, expenditure per pupil, and state GDP, all else equal, with coefficients easier to dissect and digest.

The estimated model (5) is as follows:

$$\mathbf{logIND2019 = 7.45 + .006(COL2019) - 0.006(HS2019) + 0.34(logEXDPUP2017)}$$
$$\mathbf{-0.001\ (MAN2019) + 0.02(logSTATEGDP2019)}$$

This estimated model shows that a one percent increase in expenditure per pupil leads to a 0.34% increase in per-capita income, all else equal. In addition, a one percent increase in state GDP leads to a 0.02% increase in per-capita income. The coefficients attached to these variables in other models were small and did not clearly explain the relationship to individual income. Taking the log of these variables made the correlation and impact easier to see and understand.

V. Conclusion

The goal of this research was to further expose the foundational relationship between per-capita income and college attainment. This goal was achieved as the original hypothesis was successively supported by the estimated regression models. In each regression model, college attainment was found to have a positive correlation with per-capita income and statistically significant in the estimated regression models. But it should be noted that even though in the simple linear regression model, college attainment was found to be statistically significant and the R^2 value was quite high for a simple regression, there are more variables that do indeed correlate with college attainment and do share a part in explaining the variation in per-capita income. While this research was started with the intent of model 2 being the main model to explain the variation in per-capita income, after the conduction of the first F-test, it is better to conclude that the variables HS2019 and MAN2019 should be dropped from

model 2 to form model 3 as it is a more suitable model to analyze the ceteris paribus effect of college attainment on individual income.

In acknowledgment to the some limitations faced upon the conduction of this research, model 4 was estimated, with the new variables, CST2020 and URBAN2010 being included. With the addition of these variables, the highest R^2 value was achieved with URBAN2010 reigning to be statistically significant at the 5% level though it should be noted that its adjusted R-squared value was only 0.86. It would have been beneficial to add this variable into model 3, but with the lack of recent data in the percentage of the population that resides in urban areas, it would have been too big of an assumption to make that the difference in years would have had no impact on the data. Perhaps due to these types of limitations, another sort of data might be more beneficial to discover the ceteris paribus effect of college attainment on per-capita income.

Though out of the scope of this research, the further analysis of the relationship between college attainment and per-capita income is a great tool to to utilize foundations in order to explore more complex relationships with more variables that might correlate with college attainment and inconsequence have an impact on with per-capita income. Issues such as inequality whether it pertains to income, education, or poverty can all be explored using this foundational relationship and the assumptions made from it.

Though basic in its scope, this research immensely supported the relationship between education and income further adding to the story of how closely correlated education and wage are to each other.

Appendix A. STATA Regression Model Outputs

Model 1.

```
. reg IND2019 COL2019
```

Source	SS	df	MS	Number of obs	=	51
				F(1, 49)	=	116.17
Model	3.1550e+09	1	3.1550e+09	Prob > F	=	0.0000
Residual	1.3308e+09	49	27159029	R-squared	=	0.7033
				Adj R-squared	=	0.6973
Total	4.4858e+09	50	89715022.7	Root MSE	=	5211.4

IND2019	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
COL2019	1234.615	114.5492	10.78	0.000	1004.42	1464.81
_cons	15671.61	3703.589	4.23	0.000	8228.968	23114.25

Model 2.

```
.
.
. reg IND2019 COL2019 HS2019 EXDPUP2017 MAN2019 STATEGDP2019
```

Source	SS	df	MS	Number of obs	=	51
				F(5, 45)	=	58.71
Model	3.8895e+09	5	777899141	Prob > F	=	0.0000
Residual	596255429	45	13250120.6	R-squared	=	0.8671
				Adj R-squared	=	0.8523
Total	4.4858e+09	50	89715022.7	Root MSE	=	3640.1

IND2019	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
COL2019	519.654	181.309	2.87	0.006	154.479	884.8291
HS2019	-362.0274	240.0271	-1.51	0.138	-845.4669	121.412
EXDPUP2017	1.248155	.2100921	5.94	0.000	.8250076	1.671302
MAN2019	-87.18258	164.4555	-0.53	0.599	-418.413	244.0478
STATEGDP2019	.0028094	.0011317	2.48	0.017	.00053	.0050887
_cons	31718.89	10030.16	3.16	0.003	11517.1	51920.67

```
.
```

Model 3.

```
. reg IND2019 COL2019 EXDPUP2017 STATEGDP2019
```

Source	SS	df	MS	Number of obs	=	51
				F(3, 47)	=	93.56
Model	3.8423e+09	3	1.2808e+09	Prob > F	=	0.0000
Residual	643424664	47	13689886.5	R-squared	=	0.8566
				Adj R-squared	=	0.8474
Total	4.4858e+09	50	89715022.7	Root MSE	=	3700

IND2019	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
COL2019	764.8623	106.4009	7.19	0.000	550.8114	978.9133
EXDPUP2017	1.097039	.1696752	6.47	0.000	.755697	1.438382
STATEGDP2019	.0034096	.0010819	3.15	0.003	.001233	.0055861
_cons	14831.71	2632.128	5.63	0.000	9536.55	20126.87

Model 4.

```
. reg IND2019 COL2019 HS2019 EXDPUP2017 MAN2019 STATEGDP2019 CSTLIVING2020 URBAN2010
```

Source	SS	df	MS	Number of obs	=	51
				F(7, 43)	=	45.17
Model	3.9488e+09	7	564107781	Prob > F	=	0.0000
Residual	536996667	43	12488294.6	R-squared	=	0.8803
				Adj R-squared	=	0.8608
Total	4.4858e+09	50	89715022.7	Root MSE	=	3533.9

IND2019	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
COL2019	465.1486	179.7073	2.59	0.013	102.7342	827.563
HS2019	-270.798	242.2649	-1.12	0.270	-759.3718	217.7757
EXDPUP2017	1.277209	.2304896	5.54	0.000	.8123827	1.742036
MAN2019	17.30918	171.093	0.10	0.920	-327.7328	362.3511
STATEGDP2019	.0018496	.001184	1.56	0.126	-.000538	.0042373
CSTLIVING2020	4.850126	34.702	0.14	0.889	-65.13312	74.83337
URBAN2010	98.87584	46.94788	2.11	0.041	4.19641	193.5553
_cons	22134.52	11400.46	1.94	0.059	-856.6992	45125.74

.

Model 5.

```
. reg logIND2019 COL2019 HS2019 logEXDPUP2017 MAN2019 logSTATEGDP2019
```

Source	SS	df	MS	Number of obs	=	51
				F(5, 45)	=	55.39
Model	1.16532255	5	.233064509	Prob > F	=	0.0000
Residual	.189346347	45	.004207697	R-squared	=	0.8602
				Adj R-squared	=	0.8447
Total	1.35466889	50	.027093378	Root MSE	=	.06487

logIND2019	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
COL2019	.0064682	.0030967	2.09	0.042	.000231	.0127054
HS2019	-.0093595	.0041008	-2.28	0.027	-.017619	-.0011
logEXDPUP2017	.3362319	.050123	6.71	0.000	.235279	.4371848
MAN2019	-.0026399	.0029561	-0.89	0.377	-.0085939	.003314
logSTATEGDP2019	.0249636	.0093651	2.67	0.011	.0061013	.043826
_cons	7.495397	.4057564	18.47	0.000	6.678162	8.312633

Model 6.

```
. reg logIND2019 COL2019 HS2019 logEXDPUP2017 logSTATEGDP2019
```

Source	SS	df	MS	Number of obs	=	51
				F(4, 46)	=	69.34
Model	1.1619669	4	.290491726	Prob > F	=	0.0000
Residual	.192701988	46	.004189174	R-squared	=	0.8577
				Adj R-squared	=	0.8454
Total	1.35466889	50	.027093378	Root MSE	=	.06472

logIND2019	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
COL2019	.0061532	.0030698	2.00	0.051	-.000026	.0123324
HS2019	-.0106399	.0038336	-2.78	0.008	-.0183565	-.0029233
logEXDPUP2017	.3489348	.0479563	7.28	0.000	.2524038	.4454659
logSTATEGDP2019	.0231527	.0091228	2.54	0.015	.0047895	.041516
_cons	7.420996	.3962361	18.73	0.000	6.623414	8.218578

Appendix B. Correlation Results

```
. correlate COL2019 HS2019 EXDPUP2017 MAN2019 GDP2019
(obs=51)
```

	COL2019	HS2019	EXD~2017	MAN2019	GDP2019
COL2019	1.0000				
HS2019	-0.7336	1.0000			
EXDPUP2017	0.6372	-0.1957	1.0000		
MAN2019	-0.3535	0.3825	-0.3474	1.0000	
GDP2019	0.1341	-0.2992	0.0554	-0.0742	1.0000

References

1. 11 Facts about the History of Education in America. 17 Oct. 2019, www.americanboard.org/blog/11-facts-about-the-history-of-education-in-america/.
2. "2019 United States Manufacturing Facts". *National Association of Manufacturers.* https://www.nam.org/state-manufacturing-data/2019-united-states-manufacturing-facts/#:~:text=Manufacturers%20in%20the%20United%20States,employing%208.51%25%20of%20the%20workforce.
3. "Completion Rates, 2015-2019, (Completing College, at least 25 and Older)". *U.S Department of Agricultural Services.* https://data.ers.usda.gov/reports.aspx?ID=17829
4. "Cost of Living Data Series". *Missouri Economic Research and Information.* https://meric.mo.gov/data/cost-living-data-series
5. De Gregorio, José, and Chong-hwa Yi. Education and income distribution: new evidence from cross-country data. No. 55. Centro de Economía Aplicada, Universidad de Chile, 1999.
6. Gelbrich, Judy. AMERICAN EDUCATION. Oregon State University, oregonstate.edu/instruct/ed416/ae4.html#:~:text=Compulsory%20school%20attendance%20laws%20were,of%20this%20law%20in%20place.
7. Gershon , Livia. "Where American Public Schools Came From." JSTOR Daily, daily.jstor.org/where-american-public-schools-came-from/.
8. Goldin, Claudia, and Lawrence F. Katz. "Mass secondary schooling and the state: the role of state compulsion in the high school movement." Understanding long-run economic growth: Geography, institutions, and the knowledge economy. University of Chicago Press, 2008. 275-310.
9. "Gross Domestic Product (GDP) summary, annual by state". *Bureau of Economic Analysis.* https://apps.bea.gov/itable/iTable.cfm?ReqID=70&step=1
10. Houthakker, H. S. "Education and Income." The Review of Economics and Statistics, vol. 41, no. 1, 1959, pp. 24–28. JSTOR, www.jstor.org/stable/1925454.
11. Scott A. Wolla and Jessica Sullivan, "Education, Income, and Wealth," Page One Economics®, January 2017
12. "State Annual Personal Income, 2019 (Preliminary)and State Quarterly Personal Income, 4th Quarter 2019". *Bureau of Economic Anaylsis.*

https://www.bea.gov/sites/default/files/2020-03/spi0320_0_0_0_0_0_0_0.pdf

13. "Summary of expenditures for public elementary and secondary education and other related programs". *National Center for Education Statistics.* https://nces.ed.gov/programs/digest/d19/tables/dt19_236.10.asp

14. Turner, Chad, et al. "Education and Income of the States of the United States: 1840–2000." Journal of Economic Growth 12.2 (2007): 101-158.

15. "Urban Percentage of the Population for States". *U.S. Census Bureau.* https://www.icip.iastate.edu/tables/population/urban-pct-states